Section 4.0 Baseline Sampling Duration and Frequency

4.1 Power and Sample Size

Power, in the context of this discussion, quantifies the probability that a particular statistical test and sample size will indicate that the mean or median loading has increased over the baseline, given that it truly has increased some specified amount (see Table 4.1a). The test is designed to guard against incorrectly concluding that the mean or median has increased by setting alpha at a low value. The probability is *less than* or equal to α that a statistical test and sample size will incorrectly indicate that the mean or median loading has increased over the baseline, given that it has *not* increased. If there has been a decrease in loadings, the risk of such an incorrect decision will be considerably less than alpha.

EPA evaluated the power of the statistical triggers by simulating a 60-month monitoring program for 5000 discharges, and recording the frequency with which the triggers indicated that the remining loadings exceeded baseline (see Section 3.1). The evaluations led to a choice of statistical procedures that achieve acceptable power and a reasonable balance between rates of false alarms and correct alarms.

The error rates of statistical decision procedures will depend upon the number of measurements ("sample size") used. If the false positive rate (alpha) is held constant, the power (the ability to detect an increase in pollutant load) will necessarily decrease as sample size decreases. EPA's evaluation assumed monthly sampling, using twelve samples taken over one year to characterize the baseline level, and using twelve samples taken over each year to monitor pollutant loads during remining. The performance of the evaluated statistical procedures was shown to be just adequate to meet the detailed objectives set out in Section 3.1 (see also U.S.E.P.A., 2001c) when based upon measurements taken once a month. Therefore, if these procedures are applied to measurements taken less frequently than once a month, or are applied to fewer than twelve

measurements per year (for annual triggers), the ability to detect an increase in pollutant load will necessarily be lower than intended.

4.2 Necessary Duration and Frequency of Sampling

Without an adequate duration and frequency of sampling, the statistical procedures could establish baseline levels that are either too low or too high. Baseline sample collection requirements protect both the remining operator and the environment. If baseline characterization of pre-existing pollutant discharges is inadequate (for example, if it is based on too few samples), there is a chance that an operator could consistently face noncompliance by discharging pollutant loadings above an underestimated baseline. In addition, there is the chance that environmental improvement could be jeopardized by allowing for pollutant loading discharges at high levels that still fall below an overestimated baseline. EPA believes that 12 monthly samples are the minimum to derive a statistically sound estimate of baseline (U.S.E.P.A. 2001b).

EPA has determined that the smallest acceptable number and frequency of samples is 12 monthly samples, taken consecutively over the course of one year. Twelve samples may provide less than the required power if autocorrelation is very high, if sampling duration is less than a year, or if the sampling interval is shortened (e.g., to one week) while the number of samples is not increased above 12. Therefore, EPA has required a minimum of 12 monthly samples to establish baseline.

One of the criteria for sample size is the ability to detect a change of one standard deviation above baseline loadings with reasonably high power. Discharge flows, concentrations, and loadings vary remarkably among monthly or weekly samples over the course of 1-4 years (Brady et al., 1998; EPA, 2001b). Sample coefficients of variation (CV, the ratio of standard deviation to mean) for iron loadings range from 0.62 to 2.7 for 80% of discharges (U.S.E.P.A., 2001d). Sample CVs for manganese loadings ranged from 0.54 to 1.7 for 80% of discharges (U.S.E.P.A.,

2001d). The median CV is about 1.0, thus the standard deviation is as large as the mean. Assuming that the CV remains constant at 1.0 from baseline to post-baseline, an increase of one standard deviation above baseline means that the mean loading has doubled. Thus, it is important to have a sampling frequency and duration that will permit the statistical procedures to detect increases in loadings with high probability when the standard deviations increase.

A permitting authority may want to consider requiring more than twelve samples per year during and after the baseline year in order to increase power and in order to provide a fair chance of observing a representative sample of discharge flows and loadings.

It is possible that one year of sampling may not adequately characterize baseline pollutant levels, because discharge flows can vary among years in response to inter-year variations in rainfall and ground water flow. There is some risk that the particular year chosen to characterize baseline flows and loadings will be a year of atypically high or low flow or loadings. Permitting authorities should be aware of this risk and may want to inform permittees of this risk in order to encourage multi-year characterization of baseline. To design a procedure to evaluate inter-year variations, EPA evaluated correlations between discharge flow and various parameters of existing mine discharge data and indices for which data spanning over many years are available to the public (i.e., Palmer Indices, Standardized Precipitation Index, Crop Moisture Index, Surface Water Supply Index, and USGS Current and Historical Daily Streamflow). EPA concluded that historical stream flow data from a USGS gage station associated with a discharge could be used to test whether the given baseline year was significantly different from the previous years. This would be done by comparing the mean stream flow for the baseline year to the 2.5^{th} and 97.5^{th} percentiles of annual mean stream flows prior to the baseline year. If the mean stream flow for the baseline year falls below the 2.5^{th} percentile or above the 97.5^{th} percentile, the year may have unusually low or high flow, respectively. In such cases, it may be best to continue baseline sampling for another year.

A sampling plan should be designed to prevent biased sampling. Sampling, during and after baseline, should systematically cover all periods of the year during which substantially high or

low discharge flows can be expected. Unequal sampling of months could bias the baseline mean or median toward high or low loadings by over sampling of high-flow or low-flow months. However, unequal sampling of different time periods can be accounted for by using statistical estimation procedures appropriate to stratified sampling. Stratified seasonal sampling, possibly with unequal sampling of different time periods, is a suitable alternative to regular monthly sampling, provided that correct statistical estimation procedures for stratified sampling are applied to estimate the mean, median, variance, interquartile range, and other quantities used in the statistical procedures, and provided that at least one sample is taken per month over the course of one year.

Flow measurement methods also should accurately measure flows during high-flow events. If the discharge overflows or bypasses the weir or flume, or if a measurement is not made as scheduled on a high-flow day, statistical characterizations of flow and loading will be inaccurate. The sampling location and methods should be designed as much as possible to permit access and sampling on all scheduled days, and to avoid the need to reschedule sampling because flow is extremely high.

References

Brady, K.B.C., R.J. Hornberger, and G. Fleeger, 1998. Influence of Geology on Postmining Water Quality: Northern Appalachian Basein. Chapter 8 in Coal Mine Drainage Prediction abnd Pollution Prevention in Pennsylvania. Edited by K.B.C. Brady, M.W. Smith, and J. Schueck, Department of Environmental Protection, pp. 8-1 to 8-92.

Griffiths, J.C., 1990. Letter to M. Smith, Pennsylvania Department of Environmental Resources dated January 14, 1990, 3 pages, with attachment of 54 pages titled “Chapter 3, Statistical Analysis of Mine Drainage Data”.

Hawkins, J.W., 1994. Statistical Characteristics of Coal Mine Discharges on Western Pennsylvania Remining Sites. Water Resources Bulletin, Vol. 30, No. 5, pp. 861-869.

McGill, R., J.W. Tukey, and W.A. Larsen, 1978. Variations of Box Plots. The American Scientist, Vol. 32, No. 1, pp. 12-16.

Millard, S.P., 1998. Environmental Stats for S-Plus: Users Manual for Windows and Unix. Springer-Verlag: New York, NY.

Sanders T.G., R.C. Ward, J.C. Loftis, T.D. Steele, D.D. Adrian, and V. Yevjevich, 1983. Design of Networks for Monitoring Water Quality. Water Resources Publications, Littleton, CO.

U.S. EPA, 2001a. Coal Remining Best Management Practices Guidance Manual. EPA-821-B-01-010.

U.S.E.P.A., 2001b. *Statistical Analysis of Abandoned Mine Drainage in the Assessment of Pollution Load*. EPA-821-B-01-014.

U.S.E.P.A., 2001c. "Evaluation of Statistical Triggers," memorandum in the rulemaking record (docket number DCN 3051).

U.S.E.P.A., 2001d. "Distribution & Variability of Coal Mine Discharge Loadings," memorandum in the rulemaking record (docket number DCN 3049).